Analysis of Morph-Based Speech Recognition and the Modeling of Out-of-Vocabulary Words Across Languages

Mathias Creutz[*], **Teemu Hirsimäki**[*], **Mikko Kurimo**[*], **Antti Puurula**[*], **Janne Pylkkönen**[*], **Vesa Siivola**[*], **Matti Varjokallio**[*], **Ebru Arısoy**[†], **Murat Saraçlar**[†], and **Andreas Stolcke**[‡]

[*] Helsinki University of Technology, <firstname>.<lastname>@tkk.fi,

[†] Boğaziçi University, arisoyeb@boun.edu.tr, murat.saraclar@boun.edu.tr,

[‡] SRI International / International Computer Science Institute, stolcke@speech.sri.com

Abstract

We analyze subword-based language models (LMs) in large-vocabulary continuous speech recognition across four "morphologically rich" languages: Finnish, Estonian, Turkish, and Egyptian Colloquial Arabic. By estimating n-gram LMs over sequences of $morphs$ instead of words, better vocabulary coverage and reduced data sparsity is obtained. Standard word LMs suffer from high out-of-vocabulary (OOV) rates, whereas the morph LMs can recognize previously unseen word forms by concatenating morphs. We show that the morph LMs generally outperform the word LMs and that they perform fairly well on OOVs without compromising the accuracy obtained for in-vocabulary words.

1 Introduction

As automatic speech recognition systems are being developed for an increasing number of languages, there is growing interest in language modeling approaches that are suitable for so-called "morphologically rich" languages. In these languages, the number of possible word forms is very large because of many productive morphological processes; words are formed through extensive use of, e.g., inflection, derivation and compounding (such as the English words 'rooms', 'roomy', 'bedroom', which all stem from the noun 'room').

For some languages, language modeling based on surface forms of words has proven successful, or at least satisfactory. The most studied language, English, is not characterized by a multitude of word forms. Thus, the recognition vocabulary can simply consist of a list of words observed in the training text, and n-gram language models (LMs) are estimated over word sequences. The applicability of the word-based approach to morphologically richer languages has been questioned. In highly compounding languages, such as the Germanic languages German, Dutch and Swedish, decomposition of compound words can be carried out to reduce the vocabulary size. Highly inflecting languages are found, e.g., among the Slavic, Romance, Turkic, and Semitic language families. LMs incorporating morphological knowledge about these languages can be applied. A further challenging category comprises languages that are both highly inflecting and compounding, such as the Finno-Ugric languages Finnish and Estonian.

Morphology modeling aims to reduce the out-of-vocabulary (OOV) rate as well as data sparsity, thereby producing more effective language models. However, obtaining considerable improvements in speech recognition accuracy seems hard, as is demonstrated by the fairly meager improvements (1–4 % relative) over standard word-based models accomplished by, e.g., Berton et al. (1996), Ordelman et al. (2003), Kirchhoff et al. (2006), Whittaker and Woodland (2000), Kwon and Park (2003), and Shafran and Hall (2006) for Dutch, Arabic, English, Korean, and Czech, or even the worse performance reported by Larson et al. (2000) for German and Byrne et al. (2001) for Czech. Nevertheless, clear improvements over a word baseline have been achieved for Serbo-Croatian (Geutner et al., 1998), Finnish, Estonian (Kurimo et al., 2006b) and Turkish (Kurimo et al., 2006a).

In this paper, subword language models in the recognition of speech of four languages are ana-

Proceedings of NAACL HLT 2007, pages 380–387,
Rochester, NY, April 2007.

lyzed: Finnish, Estonian, Turkish, and the dialect of Arabic spoken in Egypt, Egyptian Colloquial Arabic (ECA). All these languages are considered "morphologically rich", but the benefits of using subword-based LMs differ across languages. We attempt to discover explanations for these differences. In particular, the focus is on the analysis of OOVs: A perceived strength of subword models, when contrasted with word models, is that subword models can generalize to previously unseen word forms by recognizing them as sequences of shorter familiar word fragments.

2 Morfessor

Morfessor is an unsupervised, data-driven, method for the segmentation of words into morpheme-like units. The general idea is to discover as compact a description of the input text corpus as possible. Substrings occurring frequently enough in several different word forms are proposed as *morphs*, and the words in the corpus are then represented as a concatenation of morphs, e.g., 'hand, hand+s, left+hand+ed, hand+ful'. Through maximum a posteriori optimization (MAP), an optimal balance is sought between the compactness of the inventory of morphs, i.e., the *morph lexicon*, versus the compactness of the representation of the corpus.

Among others, de Marcken (1996), Brent (1999), Goldsmith (2001), Creutz and Lagus (2002), and Creutz (2006) have shown that models based on the above approach produce segmentations that resemble linguistic morpheme segmentations, when formulated mathematically in a probabilistic framework or equivalently using the Minimum Description Length (MDL) principle (Rissanen, 1989). Similarly, Goldwater et al. (2006) use a hierarchical Dirichlet model in combination with morph bigram probabilities.

The Morfessor model has been developed over the years, and different model versions exist. The model used in the speech recognition experiments of the current paper is the original, so-called *Morfessor Baseline* algorithm, which is publicly available for download.[1]. The mathematics of the Morfessor Baseline model is briefly outlined in the following; consult Creutz (2006) for details.

[1] http://www.cis.hut.fi/projects/morpho/

2.1 MAP Optimization Criterion

In slightly simplified form, the optimization criterion utilized in the model corresponds to the maximization of the following posterior probability:

$$P(\text{lexicon} \mid \text{corpus}) \propto$$
$$P(\text{lexicon}) \cdot P(\text{corpus} \mid \text{lexicon}) =$$
$$\prod_{\text{letters } \alpha} P(\alpha) \cdot \prod_{\text{morphs } \mu} P(\mu). \qquad (1)$$

The lexicon consists of all distinct morphs spelled out; this forms a long string of letters α, in which each morph is separated from the next morph using a morph boundary character. The probability of the lexicon is the product of the probability of each letter in this string. Analogously, the corpus is represented as a sequence of morphs, which corresponds to a particular segmentation of the words in the corpus. The probability of this segmentation equals the product of the probability of each morph token μ. Letter and morph probabilities are maximum likelihood estimates (empirical Bayes).

2.2 From Morphs to n-Grams

As a result of the probabilistic (or MDL) approach, the morph inventory discovered by the Morfessor Baseline algorithm is larger the more training data there is. In some speech recognition experiments, however, it has been desirable to restrict the size of the morph inventory. This has been achieved by setting a frequency threshold on the words on which Morfessor is trained, such that the rarest words will not affect the learning process. Nonetheless, the rarest words can be split into morphs in accordance with the model learned, by using the Viterbi algorithm to select the most likely segmentation. The process is depicted in Figure 1.

2.3 Grapheme-to-Phoneme Mapping

The mapping between graphemes (letters) and phonemes is straightforward in the languages studied in the current paper. More or less, there is a one-to-one correspondence between letters and phonemes. That is, the spelling of a word indicates the pronunciation of the word, and when splitting the word into parts, the pronunciation of the parts in isolation does not differ much from the pronunciation of the parts in context. However, a few exceptions

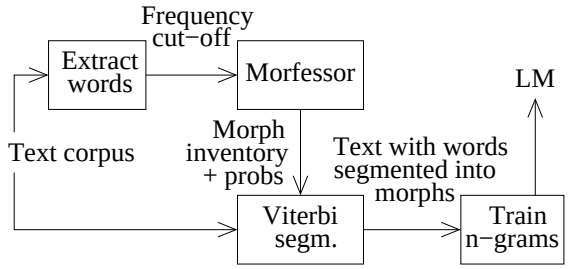

Figure 1: How to train a segmentation model using the Morfessor Baseline algorithm, and how to further train an n-gram model based on morphs.

have been treated more rigorously in the Arabic experiments: e.g., in some contexts the same (spelled) morph can have multiple possible pronunciations.

3 Experiments and Analysis

The goal of the conducted experiments is to compare n-gram language models based on morphs to standard word n-gram models in automatic speech recognition across languages.

3.1 Data Sets and Recognition Systems

The results from eight different tests have been analyzed. Some central properties of the test configurations are shown in Table 1. The Finnish, Estonian, and Turkish test configurations are slight variations of experiments reported earlier in Hirsimäki et al. (2006) (Fin1: 'News task', Fin2: 'Book task'), Kurimo et al. (2006a) (Fin3, Tur1), and Kurimo et al. (2006b) (Fin4, Est, Tur2).

Three different recognition platforms have been used, all of which are state-of-the-art large vocabulary continuous speech recognition (LVCSR) systems. The Finnish and Estonian experiments have been run on the HUT speech recognition system developed at Helsinki University of Technology.

The Turkish tests were performed using the AT&T decoder (Mohri and Riley, 2002); the acoustic features were produced using the HTK front end (Young et al., 2002). The experiments on Egyptian Colloquial Arabic (ECA) were carried out using the SRI Decipher^{TM} speech recognition system.

3.1.1 Speech Data and Acoustic Models

The type and amount of speech data vary from one language to another. The Finnish data consists of news broadcasts read by one single female speaker (Fin1), as well as an audio book read by another female speaker (Fin2, Fin3, Fin4). The Finnish acoustic models are speaker dependent (SD). Monophones (mon) were used in the earlier experiments (Fin1, Fin2), but these were later replaced by cross-context triphones (tri).

The Estonian speech data has been collected from a large number of speakers and consists of sentences from newspapers as well as names and digits read aloud. The acoustic models are speaker-independent triphones (SI tri) adapted online using Cepstral Mean Subtraction and Constrained Maximum Likelihood Linear Regression. Also the Turkish acoustic training data contains speech from hundreds of speakers. The test set is composed of newspaper text read by one female speaker. Speaker-independent triphones are used as acoustic models.

The Finnish, Estonian, and Turkish data sets contain planned speech, i.e., written text read aloud. By contrast, the Arabic data consists of transcribed spontaneous telephone conversations,[2] which are characterized by disfluencies and by the presence of "non-speech", such as laugh and cough sounds. There are multiple speakers in the Arabic data, and online speaker adaptation has been performed.

3.1.2 Text Data and Language Models

The n-gram language models are trained using the SRILM toolkit (Stolcke, 2002) (Fin1, Fin2, Tur1, Tur2, ECA) or similar software developed at HUT (Siivola and Pellom, 2005) (Fin3, Fin4, Est). All models utilize the Modified Interpolated Kneser-Ney smoothing technique (Chen and Goodman, 1999). The Arabic LM is trained on the same corpus that is used for acoustic training. This data set is regrettably small (160 000 words), but it matches the test set well in style, as it consists of transcribed spontaneous speech. The LM training corpora used for the other languages contain fairly large amounts of mainly news and book texts and conceivably match the style of the test data well.

In the morph-based models, words are split into morphs using Morfessor, and statistics are collected for morph n-grams. As the desired output of the

[2]LDC CallHome corpus of Egyptian Colloquial Arabic: `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC97S45`

Table 1: Test configurations

	Fin1	**Fin2**	**Fin3**	**Fin4**	**Est**	**Tur1**	**Tur2**	**ECA**
Recognizer	HUT	HUT	HUT	HUT	HUT	AT&T	AT&T	SRI
Speech data								
Type of speech	read	read	read	read	read	read	read	spont.
Training set [kwords]	20	49	49	49	790	230	110	160
Speakers in training set	1	1	1	1	1300	550	250	310
Test set [kwords]	4.3	1.9	1.9	1.9	3.7	7.0	7.0	16
Speakers in test set	1	1	1	1	50	1	1	57
Text data								
LM training set [Mwords]	36	36	32	150	53	17	27	0.16
Models								
Acoustic models	SD mon	SD mon	SD tri	SD tri	SI tri	SI tri	SI tri	SI tri
Morph lexicon [kmorphs]	66	66	120	25	37	52	34	6.1
Word lexicon [kwords]	410	410	410	–	60	120	50	18
Out-of-vocabulary words								
OOV LM training set [%]	5.0	5.0	5.9	–	14	5.3	9.6	0.61
OOV test set [%]	5.0	7.2	7.3	–	19	5.5	12	9.9
New words in test set [%]	2.7	3.0	3.1	1.5	3.4	1.6	1.5	9.8

speech recognizer is a sequence of words rather than morphs, the LM explicitly models word breaks as special symbols occurring in the morph sequence.

For comparison, word n-gram models have been tested. The vocabulary cannot typically include every word form occurring in the training set (because of the large number of different words), so the most frequent words are given priority; the actual lexicon sizes used in each experiment are shown in Table 1. Any word not contained in the lexicon is replaced by a special out-of-vocabulary symbol.

As words and morphs are units of different length, their optimal performance may occur at different orders of the n-gram. The best order of the n-gram has been optimized on development test sets in the following cases: Fin1, Fin2, Tur1, ECA (4-grams for both morphs and words) and Tur2 (5-grams for morphs, 3-grams for words). The models have additionally been pruned using entropy-based pruning (Tur1, Tur2, ECA) (Stolcke, 1998). In the other experiments (Fin3, Fin4, Est), no fixed maximum value of n was selected. n-Gram growing was performed (Siivola and Pellom, 2005), such that those n-grams that maximize the training set likelihood are gradually added to the model. The unrestricted growth of the model is counterbalanced by an MDL-type complexity term. The highest order of n-grams accepted was 7 for Finnish and 8 for Estonian.

Note that the optimization procedure is neutral with respect to morphs vs. words. Roughly the same number of parameters are allowed in the resulting LMs, but typically the morph n-gram LMs are smaller than the corresponding word n-gram LMs.

3.1.3 Out-of-Vocabulary Words

Table 1 further shows statistics on out-of-vocabulary rates in the data sets. This is relevant for the assessment of the word models, as the OOV rates define the limits of these models.

The OOV rate for the LM training set corresponds to the proportion of words replaced by the OOV symbol in the LM training data, i.e., words that were not included in the recognition vocabulary. The high OOV rates for Estonian (14 %) and Tur2 (9.6 %) indicate that the word lexicons have poor coverage of these sets. By contrast, the ECA word lexicon covers virtually the entire training set vocabulary.

Correspondingly, the test set OOV rate is the proportion of words that occur in the data sets used for running the speech recognition tests, but that are missing from the recognition lexicons. This value is thus the *minimum error* that can be obtained by the word models, or put differently, the recognizer is guaranteed to get at least this proportion of words wrong. Again, the values are very high for Estonian (19 %) and Tur2 (12 %), but also for Arabic (9.9 %) because of the insufficient amount of training data.

Finally, the figures labeled “new words in test set” denote the proportion of words in the test set that do not occur in the LM training set. Thus, these values indicate the minimum error achievable by *any* word model trained on the training sets available.

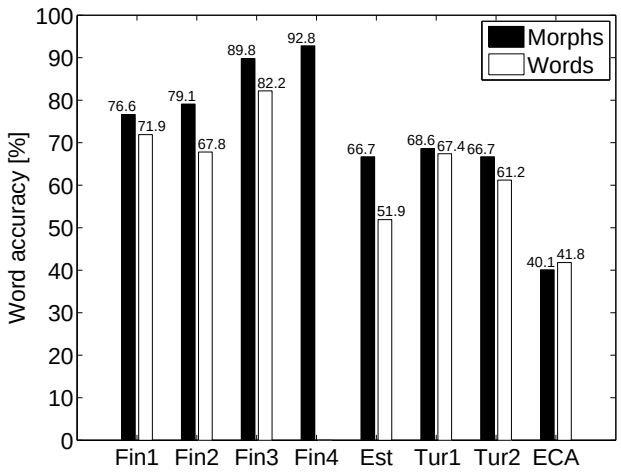

Figure 2: Word accuracies for the different speech recognition test configurations.

3.2 Results and Analysis

The morph-based and word-based results of the conducted speech recognition experiments are shown in Figure 2 (for Fin4, no comparable word experiment has been carried out). The evaluation measure used is *word accuracy* (WAC): the number of correctly recognized words minus the number of incorrectly inserted words divided by the number of words in the reference transcript. (Another frequently used measure is the *word error rate*, WER, which relates to word accuracy as WER = 100 % – WAC.)

Figure 2 shows that the morph models perform better than the word models, with the exception of the Arabic experiment (ECA), where the word model outperforms the morph model. The statistical significance of these differences is confirmed by one-tailed paired Wilcoxon signed-rank tests at the significance level of 0.05.

Overall, the best performance is observed for the Finnish data sets, which is explained by the speaker-dependent acoustic models and clean noise conditions. The Arabic setup suffers from the insufficient amount of LM training data.

3.2.1 In-Vocabulary Words

For a further investigation of the outcome of the experiments, the test sets have been partitioned into regions based on the types of words they contain. The recognition output is aligned with the reference transcript, and the regions aligned with *in-vocabulary* (IV) reference words (words contained in the vocabulary of the word model) are put in one partition and the remaining words (OOVs) are put in another partition. Word accuracies are then computed separately for the two partitions. Inserted words, i.e., words that are not aligned with any word in the reference, are put in the IV partition, unless they are adjacent to an OOV region, in which case they are put in the OOV partition.

Figure 3a shows word accuracies for the in-vocabulary words. Without exception, the accuracy for the IVs is higher than that of the entire test set vocabulary. One could imagine that the word models would do better than the morph models on the IVs, since the word models are totally focused on these words, whereas the morph models reserve modeling capacity for a much larger set of words. The word accuracies in Fig. 3a also partly seem to support this view. However, Wilcoxon signed-rank tests (level 0.05) show that the superiority of the word model is statistically significant only for Arabic and for Fin3.

With few exceptions, it is thus possible to draw the conclusion that *morph models are capable of modeling a much larger set of words than word models without, however, compromising the performance on the limited vocabulary covered by the word models in a statistically significant way*.

3.2.2 Out-of-Vocabulary Words

Since the word model and morph model perform equally well on the subset of words that are included in the lexicon of the word model, the overall superiority of the morph model needs to come from its successful coping with out-of-vocabulary words.

In Figure 3b, word accuracies have been plotted for the out-of-vocabulary words contained in the test set. It is clear that the recognition accuracy for the OOVs is much lower than the overall accuracy. Also, negative accuracy values are observed. This happens when the number of insertions exceeds the number of correctly recognized units.

In Figure 3b, if speaker-dependent and speaker-independent setups are considered separately (and Arabic is left out), there is a tendency for the morph models to recognize the OOVs more accurately, the higher the OOV rate is. One could say that a morph model has a double advantage over a corresponding word model: the larger the proportion of OOVs

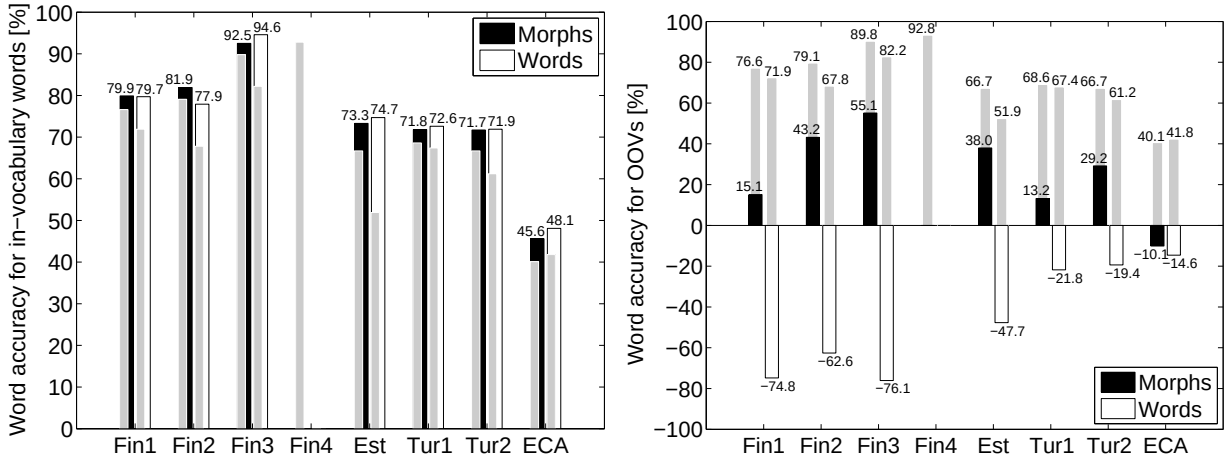

Figure 3: Word accuracies computed separately for those words in the test sets that are (a) included in and (b) excluded from the vocabularies of the word vocabulary; cf. figures listed on the row "OOV test set" in Table 1. Together these two partitions make up the entire test set vocabulary. For comparison, the results for the entire sets are shown using gray-shaded bars (also displayed in Figure 2).

in the word model is, the larger the proportion of words that the morph model can recognize but the word model cannot, a priori. In addition, the larger the proportion of OOVs, the more frequent and more "easily modelable" words are left out of the word model, and the more successfully these words are indeed learned by the morph model.

3.2.3 New Words in the Test Set

All words present in the training data (some of which are OOVs in the word models) "leave some trace" in the morph models, in the n-gram statistics that are collected for morph sequences. How, then, about new words that occur only in the test set, but not in the training set? In order to recognize such words correctly, the model must combine morphs in ways it has not observed before.

Figure 4 demonstrates that the new unseen words are very challenging. Now, also the morph models mostly obtain negative word accuracies, which means that the number of insertions adjacent to new words exceeds the number of correctly recognized new words. The best results are obtained in clean acoustic conditions (Fin2, Fin3, Fin4) with only few foreign names, which are difficult to get right using typical Finnish phoneme-to-grapheme mappings (as the negative accuracy of Fin1 suggests).

3.3 Vocabulary Growth and Arabic

Figure 5 shows the development of the size of the vocabulary (unique word forms) for growing amounts of text in different corpora. The corpora used for Finnish, Estonian, and Turkish (planned speech/text), as well as Arabic (spontaneous speech) are the LM training sets used in the experiments. Additional sources have been provided for Arabic and English: Arabic text (planned) from the FBIS corpus of Modern Standard Arabic (a collection of transcribed radio newscasts from various radio stations in the Arabic speaking world), as well as text from the New York Times magazine (English planned) and spontaneous transcribed English telephone conversations from the Fisher corpus.

The figure illustrates two points: (1) The faster the vocabulary growth is, the larger the potential advantage of morph models is in comparison to standard word models, because of OOV and data sparsity problems. The obtained speech recognition results seem to support this hypothesis; the applied morph LMs are clearly beneficial for Finnish and Estonian, mostly beneficial for Turkish, and slightly detrimental for ECA. (2) A more slowly growing vocabulary is used in spontaneous speech than in planned speech (or written text). Moreover, the Arabic 'spontaneous' curve is located fairly close

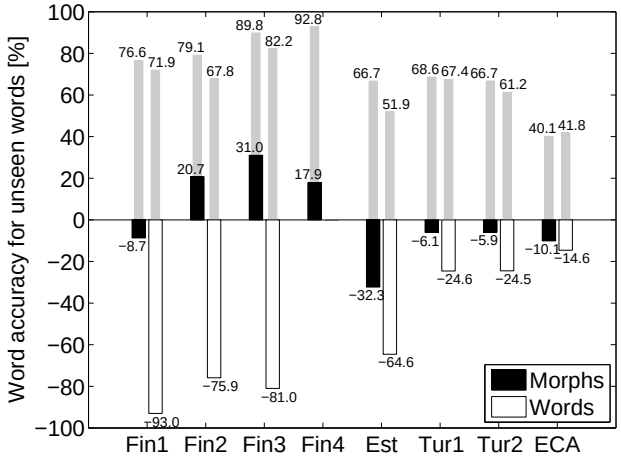

Figure 4: Word accuracies computed for the words in the test sets that do not occur at all in the training sets; cf. figures listed on the row "new words in test set" in Table 1. For comparison, the gray-shaded bars show the corresponding results for the entire test sets (also displayed in Figure 2).

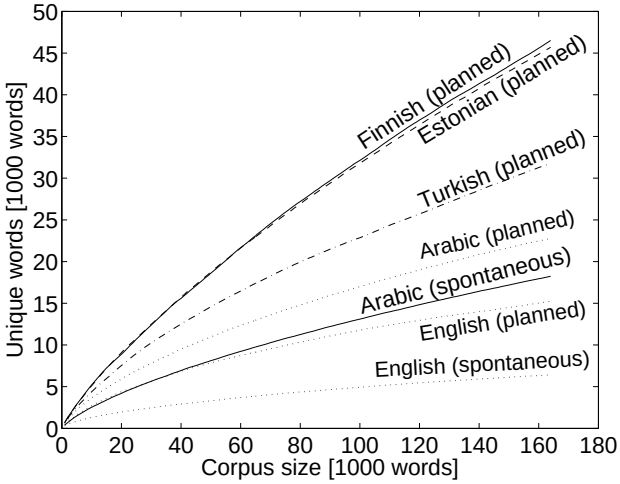

Figure 5: Vocabulary growth curves for the different corpora of spontaneous and planned speech (or written text). For growing amounts of text (word tokens) the number of unique different word forms (word types) occurring in the corpus are plotted.

to the English 'planned' curve and much below the Finnish, Estonian, and Turkish curves. Thus, even though Arabic is considered a "morphologically rich" language, this is not manifested through a considerable vocabulary growth (and high OOV rate) in the Egyptian Colloquial Arabic data used in the current speech recognition experiments. Consequently, it may not be that surprising that the morph model did not work particularly well for Arabic.

Arabic words consist of a stem surrounded by prefixes and suffixes, which are fairly successfully segmented out by Morfessor. However, Arabic also has *templatic* morphology, i.e., the stem is formed through the insertion of a vowel pattern into a "consonantal skeleton".

Additional experiments have been performed using the ECA data and Factored Language Models (FLMs) (Kirchhoff et al., 2006). The FLM is a powerful model that makes use of several sources of information, in particular a morphological lexicon of ECA. The FLM incorporates mechanisms for handling templatic morphology, but despite its sophistication, it barely outperforms the standard word model: The word accuracy of the FLM is 42.3 % and that of the word model is 41.8 %. The speech recognition implementation of both the FLM and the word model is based on *whole words* (although subword units are used for assigning probabilities to word forms in the FLM). This contrasts these models with the morph model, which splits words into subword units also in the speech recognition implementation. It seems that the splitting is a source of errors in this experimental setup with very little data available.

4 Discussion

Alternative morph-based and word-based approaches exist. We have tried some, but none of them has outperformed the described morph models for Finnish, Estonian, and Turkish, or the word and FLM models for Egyptian Arabic (in a statistically significant way). The tested models comprise more linguistically accurate morph segmentations obtained using later Morfessor versions (Categories-ML and Categories-MAP) (Creutz, 2006), as well as analyses obtained from morphological parsers.

Hybrids, i.e., word models augmented with phonemes or other subword units have been proposed (Bazzi and Glass, 2000; Galescu, 2003; Bisani and Ney, 2005). In our experiments, such models have outperformed the standard word models, but not the morph models.

Simply growing the word vocabulary to cover the

entire vocabulary of large training corpora could be one (fairly "brute-force") approach, but this is hardly feasible for languages such as Finnish. The entire Finnish LM training data of 150 million words (used in Fin4) contains more than 4 million unique word forms, a value ten times the size of the rather large word lexicon currently used. And even if a 4-million-word lexicon were to be used, the OOV rate of the test set would still be relatively high: 1.5 %.

Judging by the Arabic experiments, there seems to be some potential in Factored Language Models. The FLMs might work well also for the other languages, and in fact, to do justice to the more advanced morph models from later versions of Morfessor, FLMs or some other refined techniques may be necessary as a complement to the currently used standard n-grams.

Acknowledgments

We are most grateful to Katrin Kirchhoff and Dimitra Vergyri for their valuable help on issues related to Arabic, and to the EU AMI training program for funding part of this work. The work was also partly funded by DARPA under contract No. HR0011-06-C-0023 (approved for public release, distribution is unlimited). The views herein are those of the authors and do not necessarily reflect the views of the funding agencies.

References

I. Bazzi and J. R. Glass. 2000. Modeling out-of-vocabulary words for robust speech recognition. In *Proc. ICSLP*, Beijing, China.

A. Berton, P. Fetter, and P. Regel-Brietzmann. 1996. Compound words in large-vocabulary German speech recognition systems. In *Proc. ICSLP*, pp. 1165–1168, Philadelphia, PA, USA.

M. Bisani and H. Ney. 2005. Open vocabulary speech recognition with flat hybrid models. In *Proc. Interspeech*, Lisbon, Portugal.

M. R. Brent. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34:71–105.

W. Byrne, J. Hajič, P. Ircing, F. Jelinek, S. Khudanpur, P. Krbec, and J. Psutka. 2001. On large vocabulary continuous speech recognition of highly inflectional language — Czech. In *Proc. Eurospeech*, pp. 487–489, Aalborg, Denmark.

S. F. Chen and J. Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13:359–394.

M. Creutz and K. Lagus. 2002. Unsupervised discovery of morphemes. In *Proc. ACL SIGPHON*, pp. 21–30, Philadelphia, PA, USA.

M. Creutz. 2006. *Induction of the Morphology of Natural Language: Unsupervised Morpheme Segmentation with Application to Automatic Speech Recognition*. Ph.D. thesis, Helsinki University of Technology. `http://lib.tkk.fi/Diss/2006/isbn9512282119/`.

C. G. de Marcken. 1996. *Unsupervised Language Acquisition*. Ph.D. thesis, MIT.

L. Galescu. 2003. Recognition of out-of-vocabulary words with sub-lexical language models. In *Proc. Eurospeech*, pp. 249–252, Geneva, Switzerland.

P. Geutner, M. Finke, and P. Scheytt. 1998. Adaptive vocabularies for transcribing multilingual broadcast news. In *Proc. ICASSP*, pp. 925–928, Seattle, WA, USA.

J. Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198.

S. Goldwater, T. L. Griffiths, and M. Johnson. 2006. Contextual dependencies in unsupervised word segmentation. In *Proc. Coling/ACL*, pp. 673–680, Sydney, Australia.

T. Hirsimäki, M. Creutz, V. Siivola, M. Kurimo, S. Virpioja, and J. Pylkkönen. 2006. Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech and Language*, 20(4):515–541.

K. Kirchhoff, D. Vergyri, J. Bilmes, K. Duh, and A. Stolcke. 2006. Morphology-based language modeling for Arabic speech recognition. *Computer Speech and Language*, 20(4):589–608.

M. Kurimo, M. Creutz, M. Varjokallio, E. Arısoy, and M. Saraçlar. 2006a. Unsupervised segmentation of words into morphemes – Morpho Challenge 2005, Application to automatic speech recognition. In *Proc. Interspeech*, Pittsburgh, PA, USA.

M. Kurimo, A. Puurula, E. Arısoy, V. Siivola, T. Hirsimäki, J. Pylkkönen, T. Alumäe, and M. Saraçlar. 2006b. Unlimited vocabulary speech recognition for agglutinative languages. In *Proc. NAACL-HLT*, New York, USA.

O.-W. Kwon and J. Park. 2003. Korean large vocabulary continuous speech recognition with morpheme-based recognition units. *Speech Communication*, 39(3–4):287–300.

M. Larson, D. Willett, J. Koehler, and G. Rigoll. 2000. Compound splitting and lexical unit recombination for improved performance of a speech recognition system for German parliamentary speeches. In *Proc. ICSLP*.

M. Mohri and M. D. Riley. 2002. DCD library, Speech recognition decoder library. AT&T Labs Research. `http://www.research.att.com/sw/tools/dcd/`.

R. Ordelman, A. van Hessen, and F. de Jong. 2003. Compound decomposition in Dutch large vocabulary speech recognition. In *Proc. Eurospeech*, pp. 225–228, Geneva, Switzerland.

J. Rissanen. 1989. Stochastic complexity in statistical inquiry. *World Scientific Series in Computer Science*, 15:79–93.

I. Shafran and K. Hall. 2006. Corrective models for speech recognition of inflected languages. In *Proc. EMNLP*, Sydney, Australia.

V. Siivola and B. Pellom. 2005. Growing an n-gram model. In *Proc. Interspeech*, pp. 1309–1312, Lisbon, Portugal.

A. Stolcke. 1998. Entropy-based pruning of backoff language models. In *Proc. DARPA BNTU Workshop*, pp. 270–274, Lansdowne, VA, USA.

A. Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proc. ICSLP*, pp. 901–904. `http://www.speech.sri.com/projects/srilm/`.

E. W. D. Whittaker and P. C. Woodland. 2000. Particle-based language modelling. In *Proc. ICSLP*, pp. 170–173, Beijing, China.

S. Young, D. Ollason, V. Valtchev, and P. Woodland. 2002. *The HTK book (for version 3.2 of HTK)*. University of Cambridge.